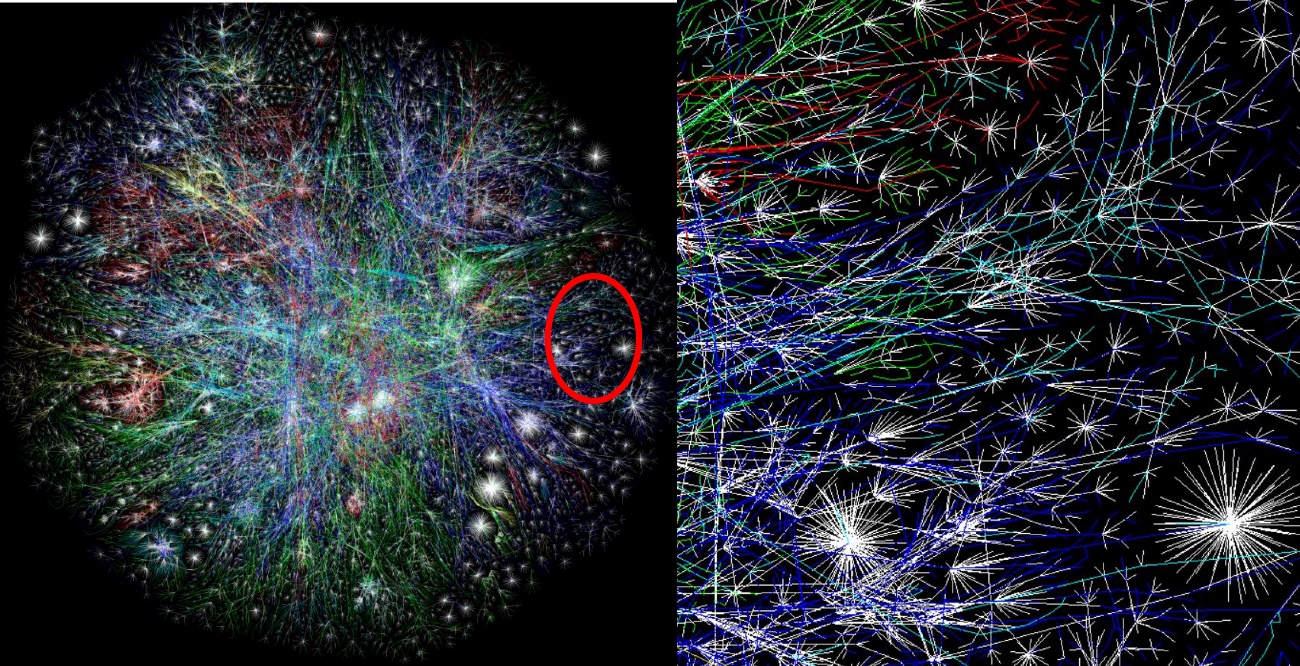
**A relação entre Data Science e Sistemas Complexos**

Talvez dois dos principais problemas em Data Science sejam um número cada vez maior de dados e como extrair informações relevantes destes. Enquanto ferramentas para a extração e armazenamento de Big Data parecem cada vez mais promissoras, compreender dados que geralmente possuem alta dimensionalidade, graus de incerteza sobre suas condições iniciais e limitantes, e heterogeneidade é um desafio considerável. Tais tipos de dados muitas vezes podem ser classificados como sistemas complexos. Porém, o que seria um sistema complexo? A definição não é única, mas podemos usar a concebida pela cientista [Melanie Mitchell](https://en.wikipedia.org/wiki/Melanie_Mitchell" \t "_blank) (2009):

*um sistema no qual grandes redes de componentes sem controle central, e regras simples de operação dão origem a um comportamento coletivo complexo, com sofisticado processamento de informações e adaptação por meio de aprendizagem ou evolução.*

Estes sistemas podem apresentar diversas características, tais como ciclos de retroalimentação positiva e negativa, não linearidade (onde o todo é diferente do que a soma das partes), comportamento caótico, e uma distribuição não gaussiana de dados, em que resultados muito distantes da média são mais prováveis do que o esperado.

Uma grande parte dos fenômenos do mundo real podem ser considerados sistemas complexos, tais como o sistema nervoso, organismos, colônias de formigas (além da própria interação entre as formigas), organizações econômicas e sociais (e.g. cidades, sistemas financeiros), ecossistemas, sistemas de infraestrutura (e.g. energia e telecomunicações), a World Wide Web, entre outros. Sistemas em que a organização central não possui um controle ou líder interno ou externo, também são conhecidos como auto-organizados. Eles são muito difíceis de serem previstos, apesar de regras e interações simples produzirem seu comportamento, conhecido em uma escala macro (referente ao todo), como emergente.



O [Projeto Opte](http://www.opte.org/maps/" \t "_blank) é uma iniciativa gratuita e de código aberto que tem como objetivo criar uma representação visual dos espaços metafísicos da Internet. Esta imagem de 2003 representada por uma rede complexa, segundo seu criador Barret Lyon, foi baseada em uma tecnologia chamada‘ traceroute ’, que na verdade passa por cada sequência da Internet e cria um traço visual. Os [nós (vértices)](https://pt.wikipedia.org/wiki/V%C3%A9rtice_(teoria_dos_grafos)" \t "_blank)representam servidores e hosts individuais, e os [links](https://pt.wikipedia.org/wiki/Aresta_(teoria_dos_grafos)" \t "_blank) (arestas) representam a fibra, o cobre ou outra conexão entre eles. Do lado esquerdo, está a imagem da rede complexa, com a elipse em vermelho em zoom do lado direito.

A black background with white dots

Description automatically generated with low confidence

Esta é uma simulação de autômato celular chamada de “[Conway’s Game of Life](https://pt.wikipedia.org/wiki/Jogo_da_vida" \t "_blank)”, em que regras simples e atividades aparentemente aleatórias podem criar padrões estruturados que parecem ter vida “inteligente”.

Um sistema pode parecer complexo, mas não ser um sistema complexo *per se*, apenas complicado. Problemas complicados podem ser de difícil resolução, com muito elementos interagindo entre si, mas se eles interagem de maneira linear, sua resolução correta é apenas questão de tempo e pode ser feita em blocos aditivos (Nason, 2017). O grande problema é aplicar abordagens simples para problemas complexos e que são inerentemente não-lineares, isto pode levar a conclusões errôneas e perigosas. Por exemplo, todas as variáveis relacionadas ao clima, tais como temperatura, umidade, pressão do ar, precipitação de chuva e etc, podem ser mensuradas corretamente, porém é muito difícil ter previsibilidade sobre o clima em uma escala de tempo maior do que dias. Isto ocorre porque o clima é um sistema caótico, onde uma pequena mudança nas condições iniciais de uma parte do sistema pode ter grandes consequências no sistema todo, tal fenômeno ficou conhecido na cultura popular como o “efeito borboleta”. O efeito borboleta foi nomeado desta forma devido a uma palestra feita por [Edward Lorenz](https://pt.wikipedia.org/wiki/Edward_Lorenz" \t "_blank) em uma conferência científica, entitulada *“O bater das asas de uma borboleta no Brasil desencadeou um tornado no Texas?”* O propósito de sua pergunta provocativa, disse ele, era ilustrar a ideia de que alguns sistemas dinâmicos complexos exibem comportamentos imprevisíveis, de modo que pequenas variações nas condições iniciais podem ter efeitos profundos e amplamente divergentes sobre os resultados do sistema. Devido à sensibilidade desses sistemas, os resultados de suas dinâmicas são imprevisíveis. Outro exemplo é o mercado financeiro, que também possui um nível de imprevisibilidade muito mais alto que outros sistemas, pode-se compreender todas as diferentes partes do sistema, mas à medida que ele evolui, elas frequentemente interagem de maneiras imprevisíveis.

A picture containing map, text, athletic game, sport

Description automatically generated

O pêndulo duplo é um bom exemplo de um sistema que se comporta de maneira extremamente diferente com apenas mudanças pequenas das condições iniciais.

Ok, é um tanto quanto evidente que o mundo real é mais complexo do que conseguimos observar, mas isso faz com que tenhamos limitações intrínsecas para o estudo da ciência de dados de certos fenômenos? Não necessariamente, pelo menos em alguns casos. Existe uma área acadêmica focada nos estudos de sistemas complexos, e suas ferramentas analíticas podem ser também utilizadas em um contexto voltado para Data Science. Estas ferramentas podem ser usadas em modelagem, para descobrir através de simulações quais padrões emergem das interações entre as variáveis. Os resultados das simulações podem gerar novos insights sobre os dados, além de criar novas variáveis e medidas que podem ser analisadas por métodos mais tradicionais. Podemos fazer um um resumo de algumas das principais abordagens utilizadas:

* Teoria de redes (i.e. Teoria dos Grafos): Talvez uma das interseções mais conhecidas entre sistemas complexos e data science. Uma rede complexa é descrita como um sistema de nós (objetos) conectados por interações (links ou arestas), que se comunicam entre si, exibindo padrões de conexão que não são puramente regulares, nem puramente aleatórios (Newman 2003). Dados de redes sociais são geralmente associados com modelagem de grafos e análises de redes complexas, tais como a análise do rank de diferentes nós, coeficientes de cluster, medidas de centralidade, entre outras, além da possibilidade do uso de redes bayesianas. Além de redes sociais, relações entre elementos culturais (e.g. religiões), tecnológicos (e.g. sistemas de transporte e comunicação), financeiros (e.g. mercados financeiros, comércio internacional) e biológicos (e.g. interações entre genes e organismos, transmissão de doenças) podem ser representadas como redes complexas.
* Modelagem baseada em agentes (i.e. Agent-based models): permite a modelagem de agentes independentes que interagem entre si, e o efeito dessas interações no sistema geral. Usando regras de interação simples, é possível observar um comportamento emergente complexo que pode fornecer percepções sobre um sistema. ABMs são usados em diversas situações, por exemplo, para estudar a dinâmica de multidões, o comportamento de organismos ou animais que agem livremente, fenômenos epidemiológicos ([como em estudos de previsão do contágio de Covid-19](https://towardsdatascience.com/agent-based-modeling-will-unleash-a-new-paradigm-of-machine-learning-ff6d3b1ac940" \t "_blank)) ou para estudar e projetar espaços físicos de acordo com os fluxos de tráfego humano e logística. ABMs são inerentemente voltados para o futuro e não requerem grandes volumes de dados históricos com alta qualidade, apenas hipóteses ou conhecimento prévio de como um sistema ou seus componentes podem funcionar.
* Autômatos celulares (AC): modelos computacionais discretos baseados em grade, nos quais elementos simples ("células") estão dispostos em uma rede regular. O estado da célula no próximo instante de tempo é determinada de acordo com o seu próprio estado atual, o das células vizinhas e de um conjunto de regras, conhecidas como regras de transições locais. Uma configuração inicial do autômato, aparentemente simples, pode produzir resultados em que a conjuntura da matemática dos estados apresentará um alto nível de complexidade (Wolfram, 1994). Aplicações de autômatos celulares em Data Science são observadas em [Deep Learning](https://towardsdatascience.com/cellular-automaton-and-deep-learning-2bf7c57139b3" \t "_blank).
* Métodos de Monte Carlo: uma abordagem que inclui vários algoritmos que tem em comum o uso da geração de números aleatórios para resolver problemas determinísticos. O uso da modelagem de MC é extenso, com contribuições na física, biologia, teoria de jogos, e finanças. Frequentemente aplicado para análise de risco e decisão, por exemplo, para examinar a probabilidade de perda em uma carteira de investimentos. Dentre os algoritmos de MC existentes, Monte Carlo tree search (MCTS) provou ser muito eficaz no tratamento de grandes volumes de dados para aplicações de inteligência artificial (Silver, 2016).

A picture containing outdoor, swimming, ocean floor

Description automatically generated

O quão fundo precisamos ir?

Determinar qual abordagem utilizar depende do problema que se está analisando, muitas vezes mais de um método pode ser usado devido a oferecer informações diferentes e complementares sobre o sistema estudado. Pretendo postar mais sobre dois tópicos aqui mencionados, que utilizei em meus estudos acadêmicos de sistemas complexos biológicos, redes complexas e ABMs. Em suma, talvez a grande questão é que sistemas complexos são….complexos! espera-se que não exista uma resposta fácil para como abordá-los, ainda mais em um contexto de predição. O que deve-se evitar é considerar um sistema complexo como simples ou que pode ser linearilizado, sem uma checagem aprofundada de como perturbações afetam o sistema. As consequências de uma previsão falha podem ser catastróficas. Incorporar o poder de ferramentas analíticas de sistemas complexos com inteligência artificial pode ser o melhor caminho para uma previsão mais refinada do futuro.

Bibliografia

Mitchell, M. (2009). *Complexity: A guided tour*. Oxford University Press.

Nason, R. (2017). *It’s Not Complicated: The Art and Science of Complexity in Business*. University of Toronto Press.

Newman, M. E. (2003). The structure and function of complex networks. *SIAM review*, *45*(2), 167–256.

Silver D, Huang A, Maddison CJ, Guez A, Sifre L. Mastering the game of go with deep neural networks and tree search. Nature 2016;529:484489.

Wolfram, S. (1994). Cellular Automata and Complexity. Collected Papers. MA Reading: Addison-Wesley. 1a edition

<https://towardsdatascience.com/cellular-automaton-and-deep-learning-2bf7c57139b3>

<https://towardsdatascience.com/agent-based-modeling-will-unleash-a-new-paradigm-of-machine-learning-ff6d3b1ac940>